To Extend or Not to Extend?
Context-driven Corpus Enrichment
Data Linking for Humanities Research – Workshop

Tanya Braun

Institute of Information Systems
University of Lübeck

October 21, 2019

A sneak preview of a paper of the same name by Kuhr, B, Bender, and Möller accepted for publication at AI 2019
Understanding Written Artefacts
Eva Wilden, Tamil Satellite Stanzas: Genres and Distribution

Transcript

Translation

Annotations for transcript or translation

- Hyperlinks (doi)
- Tags (strings)
- Descriptions of objects (entities) and relations between them
Annotations

Descriptions of objects (entities) and relations between them

Figure: Objects and their relations in a graph
Annotations

Descriptions of objects (entities) and relations between them

Figure: Objects and their relations in a graph

Long-term Goal

Figure: A probabilistic relational model allowing for modelling uncertainty
A Corpus of Documents and Annotations

General Setting

- Corpus = set of text documents

Assumption
Annotations are relevant for a given task, i.e., reflect the context.
A Corpus of Documents and Annotations

General Setting

- Corpus = set of text documents
- Each document has a set of annotations

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

<annotation1>
<annotation2>
<annotation3>
A Corpus of Documents and Annotations

General Setting

- Corpus = set of text documents
- Each document has a set of annotations
- Each annotation is associated with words at a specific location
A Corpus of Documents and Annotations

General Setting

- Corpus = set of text documents
- Each document has a set of annotations
- Each annotation is associated with words at a specific location

Assumption

Annotations are relevant for a given task, i.e., reflect the context.
A Selection of Tasks in Data Linking

Standard Document Retrieval

In a *fixed* corpus, find documents
A Selection of Tasks in Data Linking

Standard Document Retrieval

In a *fixed* corpus, find documents

• forming a cluster of documents similar in some regard,
A Selection of Tasks in Data Linking

Standard Document Retrieval

In a *fixed* corpus, find documents

- forming a cluster of documents similar in some regard,
- that fit some search terms, or
A Selection of Tasks in Data Linking

Standard Document Retrieval

In a *fixed* corpus, find documents

- forming a cluster of documents similar in some regard,
- that fit some search terms, or
- are similar to a given document.
A Selection of Tasks in Data Linking

Standard Document Retrieval

In a fixed corpus, find documents

• forming a cluster of documents similar in some regard,

• that fit some search terms, or

• are similar to a given document.

Solve the task using

• words, e.g., tf.idf,

• topics, e.g., LDA, or

• annotations (depends on type of annotations), e.g., entity matching.
A Selection of Tasks in Data Linking
Standard Document Retrieval

In a *fixed* corpus, find documents
• forming a cluster of documents similar in some regard,
• that fit some search terms, or
• are similar to a given document.

Solve the task using
• words, e.g., tf.idf,
A Selection of Tasks in Data Linking

Standard Document Retrieval

In a *fixed* corpus, find documents

- forming a cluster of documents similar in some regard,
- that fit some search terms, or
- are similar to a given document.

Solve the task using

- words, e.g., tf.idf,
- topics, e.g., LDA, or
A Selection of Tasks in Data Linking

Standard Document Retrieval

In a fixed corpus, find documents
• forming a cluster of documents similar in some regard,
• that fit some search terms, or
• are similar to a given document.

Solve the task using
• words, e.g., tf.idf,
• topics, e.g., LDA, or
• annotations (depends on type of annotations), e.g., entity matching.
Extend a corpus with a new document
A Selection of Tasks in Data Linking
Corpus Extension as a Formalisation of Recommending Documents

Extend a corpus with a **new document** only if the **document**

*provides additional data relevant for a given task, i.e., adds value in a given context.*
A Selection of Tasks in Data Linking
Corpus Extension as a Formalisation of Recommending Documents

Extend a corpus with a new document only if the document provides additional data relevant for a given task, i.e., adds value in a given context.

→ Corpus enrichment
A Selection of Tasks in Data Linking

Corpus Extension as a Formalisation of Recommending Documents

Extend a corpus with a **new document** only if the **document**

*provides additional data relevant for a given task, i.e., adds value in a given context.*

→ Corpus enrichment

Make decision based on

- words, BUT: not context-specific
Extend a corpus with a new document only if the document provides additional data relevant for a given task, i.e., adds value in a given context.

→ Corpus enrichment

Make decision based on

- words, BUT: not context-specific
- topics, BUT: possibly inconclusive
A Selection of Tasks in Data Linking
Corpus Extension as a Formalisation of Recommending Documents

Extend a corpus with a new document only if the document
provides additional data relevant for a given task, i.e., adds value in a given context.

→ Corpus enrichment

Make decision based on
• words, BUT: not context-specific
• topics, BUT: possibly inconclusive
• annotations?
Annotations for Corpus Enrichment

Foundation

Assumption

Annotations are relevant for a given task, i.e., reflect the context.
Annotations for Corpus Enrichment

Foundation

Assumption

Annotations are relevant for a given task, i.e., reflect the context.

Proposition

Annotations generate the words in a document.
Understanding Written Artefacts

Assumption

Annotations are relevant for a given task, i.e., reflect the context.

Proposition

Annotations generate the words in a document.
Annotations for Corpus Enrichment

Foundation

Proposition

Annotations generate the words in a document.
Annotations for Corpus Enrichment

Foundation

Proposition

Annotations generate the words in a document.

Question

How much of the document can we generate with high probability given the annotations in the corpus?
Decision

Based on answer to how much is generated with high probability: decide extension (IN/OUT)
Annotations for Corpus Enrichment

How does this question help to decide “To extend or not to extend?”?

Decision

Based on answer to how much is generated with high probability: decide extension (IN/OUT)

- Generate large part with high probability: OUT (→ known).

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibendum
Annotations for Corpus Enrichment

How does this question help to decide “To extend or not to extend?”?

Decision

Based on answer to how much is generated with high probability: decide extension (IN/OUT)

- Generate large part with high probability: OUT (\(\rightarrow\) known).
- Probability low: OUT (\(\rightarrow\) unrelated).
Annotations for Corpus Enrichment

How does this question help to decide “To extend or not to extend?”?

Decision

Based on answer to how much is generated with high probability: decide extension (IN/OUT)

- Generate large part with high probability: OUT (→ known).
- Probability low: OUT (→ unrelated).
- Generate only some parts with high probability: IN (→ extension).
Annotations for Corpus Enrichment

Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Corpus documents (offline): build vector representation of annotations with respect to words occurring with annotations

New document: for word chunks, build vector representation of the words occurring in the chunk

Use cosine similarity to find annotation whose vector representation is most similar to the words of a chunk:

\[ \text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \]
Annotations for Corpus Enrichment

Approach

- Corpus documents (offline): build vector representation of annotations with respect to words occurring with annotations.

\[
\begin{bmatrix}
\begin{array}{cccc}
w_1 & w_2 & w_3 & \cdots & w_n \\
\end{array}
\end{bmatrix}
\begin{bmatrix}
\begin{array}{cccc}
v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\
v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n}
\end{array}
\end{bmatrix}
\]

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend.
Annotations for Corpus Enrichment

Approach

- Corpus documents (offline): build vector representation of annotations with respect to words occurring with annotations
- New document: for word chunks, build vector representation of the words occurring in the chunk
Annotations for Corpus Enrichment

**Approach**

- Corpus documents (offline): build vector representation of annotations with respect to words occurring with annotations
- New document: for word chunks, build vector representation of the words occurring in the chunk
- Use cosine similarity to find annotation whose vector representation is most similar to the words of a chunk:

\[
\text{sim}(A, B) = \cos(\angle A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}
\]
Annotations for Corpus Enrichment

Approach

• Simplified representation of corpus annotations \( t_i \) with two words in the vocabulary
Annotations for Corpus Enrichment

Approach

- Simplified representation of corpus annotations $t_i$ with two words in the vocabulary
- Representation of vector representation of word chunk $t'$

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim, in massa tempor nec feugiat. Nunc aliquet bibend

$\langle v'_1, \ldots, v'_n \rangle$

$\langle v''_1, \ldots, v''_n \rangle$

$\langle v'''_1, \ldots, v'''_n \rangle$
Annotations for Corpus Enrichment

Approach

- Simplified representation of corpus annotations $t_i$ with two words in the vocabulary
- Representation of vector representation of word chunk $t'$
- Angle $\theta'_1$ between $t_1$ and $t'$ smallest compared to $t_2$, $t_3$
Annotations for Corpus Enrichment

Approach

- Simplified representation of corpus annotations $t_i$ with two words in the vocabulary
- Representation of vector representation of word chunk $t'$
- Angle $\theta'_1$ between $t_1$ and $t'$ smallest compared to $t_2$, $t_3$
  \[\text{→ Find } t_i \text{ with smallest angle for each word chunk}\]

$\begin{align*}
\theta'_1 & \text{ between } t_1 \text{ and } t' \text{ smallest compared to } t_2, t_3 \\
\text{→ Find } t_i \text{ with smallest angle for each word chunk}
\end{align*}$
Annotations for Corpus Enrichment

Approach

- Simplified representation of corpus annotations $t_i$ with two words in the vocabulary
- Representation of vector representation of word chunk $t'$
- Angle $\theta'_1$ between $t_1$ and $t'$ smallest compared to $t_2, t_3$
  \[ \rightarrow \text{Find } t_i \text{ with smallest angle for each word chunk} \]

Use set of $t_i$'s for all word chunks $t'$ in the new document and their similarities for decision
Understanding Written Artefacts

Possible Extension

Embed further data into corpus vector representation, e.g.,

- materials
- spectrograms

→ Cosine similarity over words and materials (, . . .)

→ Use for corpus enrichment or document retrieval
Annotations for Corpus Enrichment

Results (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

- Corpus: Wikipedia articles about cities
- New document:
  - $d_{sim}$: known
  - $d_{ext}$: extended
  - $d_{rev}$: revised
  - $d_{unrel}$: unrelated

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat.

Nunc aliquet bibendum
Annotations for Corpus Enrichment

Results (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

- Corpus: Wikipedia articles about cities
- New document:
  - \(d_{sim}\): known
  - \(d_{ext}\): extended
  - \(d_{rev}\): revised
  - \(d_{unrel}\): unrelated

\[
\begin{array}{c|c|c|c|c}
\text{Similarity} & 0.2 & 0.4 & 0.6 & 0.8 \\
\hline
d_{sim} & \text{max.sim} & \text{min.sim} & \Delta_{max,min} & \text{avg.sim} \\
d_{ext} & \text{max.sim} & \text{min.sim} & \Delta_{max,min} & \text{avg.sim} \\
d_{rev} & \text{max.sim} & \text{min.sim} & \Delta_{max,min} & \text{avg.sim} \\
d_{unrel} & \text{max.sim} & \text{min.sim} & \Delta_{max,min} & \text{avg.sim} \\
\end{array}
\]
Annotations for Corpus Enrichment

Challenges (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

Influencing factors

• Corpus size
• Quality of annotations
• Indicators

→ No single indicator to rule them all!
→ Limited transfer between corpora!
**Annotations for Corpus Enrichment**

Challenges (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

---

<table>
<thead>
<tr>
<th>Indicator $l$</th>
<th>city corpus</th>
<th>president corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$d_{sim}$ $d_{ext}$ $d_{rev}$ $d_{unrel}$</td>
<td>$d_{sim}$ $d_{ext}$ $d_{rev}$ $d_{unrel}$</td>
</tr>
<tr>
<td>Max Sim.</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Min Sim.</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>$\Delta_{max,min}$</td>
<td>$-$</td>
<td>o</td>
</tr>
<tr>
<td>Avg. Sim.</td>
<td>$+$</td>
<td>$+$</td>
</tr>
<tr>
<td>Max.$\Delta_{win}$</td>
<td>$-$</td>
<td>o</td>
</tr>
</tbody>
</table>

“+”: $l \geq 0.7$, “−”: $l \leq 0.3$, “○”: $0.3 < l < 0.7$
## Annotations for Corpus Enrichment

Challenges (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

<table>
<thead>
<tr>
<th>Indicator ( l )</th>
<th>( d_{sim} )</th>
<th>( d_{ext} )</th>
<th>( d_{rev} )</th>
<th>( d_{unrel} )</th>
<th>( d_{sim} )</th>
<th>( d_{ext} )</th>
<th>( d_{rev} )</th>
<th>( d_{unrel} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Sim.</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>Min Sim.</td>
<td>+</td>
<td>o</td>
<td>o</td>
<td>−</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>−</td>
</tr>
<tr>
<td>( \Delta_{max, min} )</td>
<td>−</td>
<td>o</td>
<td>−</td>
<td>o</td>
<td>−</td>
<td>o</td>
<td>−</td>
<td>o</td>
</tr>
<tr>
<td>Avg. Sim.</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>Max.( \Delta_{win} )</td>
<td>−</td>
<td>o</td>
<td>−</td>
<td>o</td>
<td>−</td>
<td>o</td>
<td>−</td>
<td>o</td>
</tr>
</tbody>
</table>

“+”: \( l \geq 0.7 \), “−”: \( l \leq 0.3 \), “o”: \( 0.3 < l < 0.7 \)
Annotations for Corpus Enrichment
Challenges (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

<table>
<thead>
<tr>
<th>Indicator</th>
<th>city corpus</th>
<th>president corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Sim.</td>
<td>$d_{sim}$</td>
<td>$d_{ext}$</td>
</tr>
<tr>
<td></td>
<td>$d_{rev}$</td>
<td>$d_{unrel}$</td>
</tr>
<tr>
<td>Min Sim.</td>
<td>$d_{sim}$</td>
<td>$d_{ext}$</td>
</tr>
<tr>
<td></td>
<td>$d_{rev}$</td>
<td>$d_{unrel}$</td>
</tr>
<tr>
<td>$\Delta_{max,min}$</td>
<td>$d_{sim}$</td>
<td>$d_{ext}$</td>
</tr>
<tr>
<td></td>
<td>$d_{rev}$</td>
<td>$d_{unrel}$</td>
</tr>
<tr>
<td>Avg. Sim.</td>
<td>$d_{sim}$</td>
<td>$d_{ext}$</td>
</tr>
<tr>
<td></td>
<td>$d_{rev}$</td>
<td>$d_{unrel}$</td>
</tr>
<tr>
<td>Max. $\Delta_{win}$</td>
<td>$d_{sim}$</td>
<td>$d_{ext}$</td>
</tr>
<tr>
<td></td>
<td>$d_{rev}$</td>
<td>$d_{unrel}$</td>
</tr>
</tbody>
</table>

"+": $I \geq 0.7$, "−": $I \leq 0.3$, "○": $0.3 < I < 0.7$
Annotations for Corpus Enrichment

Benefits

• Enrich corpus with documents that add value → Recommendations
• Use similarities as guideline to unknown portions
• Use annotations as a starting point for annotating new document
• Augment annotations of corpus documents with new annotations of unknown document portions
Understanding Written Artefacts
Possible Directions

Transcript

Translation

• Parse critical editions/manuscripts (and existing annotations)
  → Annotate translations with translations of referenced books (managing copyrights)
  → Add annotations to manuscript without annotations using books that reference the manuscript